Count data models with correlated unobserved heterogeneity

Boes, Stefan

Abstract: As previously argued, the correlation between included and omitted regressors generally causes inconsistency of standard estimators for count data models. Non-linear instrumental variables estimation of an exponential model under conditional moment restrictions is one of the proposed remedies. This approach is extended here by fully exploiting the model assumptions and thereby improving efficiency of the resulting estimator. Empirical likelihood in particular has favourable properties in this setting compared with the two-step generalized method of moments, as demonstrated in a Monte Carlo experiment. The proposed method is applied to the estimation of a cigarette demand function.

DOI: https://doi.org/10.1111/j.1467-9469.2010.00689.x

Scandinavian Journal of STATISTICS

Theory and Applications

Scandinavian Journal of Statistics, Vol. 37: 382–402, 2010
doi: 10.1111/j.1467-9469.2010.00689.x

Count Data Models with Correlated Unobserved Heterogeneity

STEFAN BOES

Socioeconomic Institute, University of Zurich

ABSTRACT. As previously argued, the correlation between included and omitted regressors generally causes inconsistency of standard estimators for count data models. Non-linear instrumental variables estimation of an exponential model under conditional moment restrictions is one of the proposed remedies. This approach is extended here by fully exploiting the model assumptions and thereby improving efficiency of the resulting estimator. Empirical likelihood in particular has favourable properties in this setting compared with the two-step generalized method of moments, as demonstrated in a Monte Carlo experiment. The proposed method is applied to the estimation of a cigarette demand function.

Key words: approximating functions, instrumental variables, non-parametric likelihood, optimal instruments, Poisson model, semiparametric efficiency

1. Introduction

Regression models for count data have become a standard tool in empirical work with applications in all areas of specialization. Examples include the number of patents applied for by a firm (Hausman *et al.*, 1984), the number of supreme court appointments (King, 1987), the number of epileptic seizures (Thall & Vail, 1990), the number of doctor visits (Pohlmeier & Ulrich, 1995), the number of children born to a woman (Winkelmann & Zimmermann, 1994), the number of days a worker is absent from his job (Delgado & Kniesner, 1997) and the number of cubes in a tower building test as a measure of fine motor development of children (Cheung, 2002).

The basic empirical model in most applications is the Poisson regression model. The Poisson model assumes that the count variable Y follows a Poisson distribution given a vector of observed variables X, formally $Y \mid X \sim \text{Poisson}(\mu_X)$, with log-linear specification of the intensity parameter μ_X. Although certainly being appropriate in many cases, the Poisson model may not always display the true data-generating process. For example, it presumes that the researcher is able to account for the full amount of individual heterogeneity just by including X, additional unobserved heterogeneity is not allowed for and ruled out by the model assumptions. Various generalizations have been proposed that account for such unobserved heterogeneity. The standard approaches employ mixture distributions, either parametrically by introducing, for example, Gamma-distributed heterogeneity (the negative binomial models), or semiparametrically without specifying the form of the mixing distribution (Gurmu *et al.*, 1998). Winkelmann (2008) gives an overview.

Mullahy (1997) extends the literature to the important case when independence between observed and unobserved heterogeneity fails, for example, owing to endogeneity. He considers the conditional expectation function $E(Y \mid X, v)$, specified as the exponential of a linear predictor $X'\beta$ with multiplicative unobserved heterogeneity v. Mullahy (1997) points out that, given non-zero correlation between X and v, standard estimators like Poisson pseudo-maximum likelihood (PML) or non-linear least squares will generally be inconsistent for β because the usual residual function is not orthogonal to X. Also, a non-linear instrumental

variables strategy based on this residual function will be inconsistent owing to the non-separability of X and v.

Fortunately, a simple transformation of the model yields a residual function $\rho(Y, X; \beta)$ that is additively separable in X and v, and the assumption of mean independence between the latter and the vector of instruments Z can be used to construct conditional moment restrictions $E[\rho(Y, X; \beta) \mid Z] = 0$. As proposed by Mullahy (1997), estimation can then be based on the generalized method of moments (GMM) using moment functions $g(Y, X, Z; \beta) = a(Z)\rho(Y, X; \beta)$ for some function $a(Z)$, and the resulting estimator for β will be consistent and asymptotically normally distributed. The estimator is not necessarily efficient, though, because its asymptotic variance depends on the choice of $a(Z)$.

The aim of this article is to extend Mullahy's (1997) approach using optimal instruments $a^*(Z)$ that fully utilize the information given by the conditional moment restrictions. Compared with Mullahy's work, this article makes a formal statement of how the optimal instrument matrix should be chosen. This provides an important guideline for practitioners who estimate exponential models with potentially endogenous regressors. Moreover, the article emphasizes the importance of model assumptions, in particular the assumptions on the instrument vector Z, first for the estimation procedure itself, and second for the properties of the resulting estimator. To the best of my knowledge, this has not sufficiently been considered in the previous literature.

The article proceeds as follows. The model and moment conditions will be laid out in the next section. Special attention will be given to the construction of the optimal instrument matrix $a^*(Z)$. Section 3 discusses the estimation methods, GMM and empirical likelihood (EL) estimation, in the given model context. Section 4 compares the properties of the estimators in a simulated data environment. The results indicate advantages of the EL estimator over the two-step GMM estimator in terms of (small sample) bias and efficiency. Section 5 applies the methods to estimate a cigarette demand function. Fully exploiting the model assumptions considerably improves efficiency. For example, approximating the optimal function $a^*(Z)$ for only one instrument more than doubles the t-statistic for the parameter of interest compared with the baseline instrument specification.

2. Exponential model, heterogeneity and moment conditions

Let Y denote a random variable with support being the non-negative integers, X denote a $k \times 1$ vector of explanatory variables (including a constant) and Z denote a $q \times 1$ vector of instruments ($q \geq k$) with properties to be defined next. Assume that n observations of (Y, X, Z) form a random sample of the population, and suppose that the main objective is to estimate the effect of elements of X on the conditional expectation $E(Y \mid X)$.

Specifically, the data-generating process is assumed to be consistent with the conditional expectation function

$$E(Y \mid X, v; \beta) = \exp(X'\beta)v, \tag{1}$$

where β is the $k \times 1$ vector of unknown parameters, and $v = \exp(\omega) > 0$ is an unobserved random variable. The specification of the conditional expectation function explicitly accounts for observed heterogeneity (through X) and unobserved heterogeneity (through v). Without loss of generality the normalization $E(v) = 1$ can be invoked if a constant term is included in X. Note that observable and unobservable characteristics are treated symmetrically in (1) because the conditional expectation function is log-linear in both X and ω. The specific form of the conditional expectation function might appear restrictive at first, but there is no *a priori* reason for X and ω to enter asymmetrically. Moreover, the linear index $X'\beta$ is sufficiently

flexible to approximate any non-linear function in the regressors arbitrarily close, and the exponential function ensures (1) to be positive, as required for a count-dependent variable. Strictly speaking, it is not necessary for (1) to be fulfilled that Y is a count. What follows is equally relevant to any other data-generating process consistent with such an exponential conditional expectation function. An exponential function with continuous Y was used, for example, in Mullahy (1997) where the dependent variable is the birthweight. Exponential functional forms should also be used to estimate gravity equations (Santos Silva & Tenreyro, 2006).

The specification of the conditional expectation function implies that

$$Y = \exp(X'\beta)v + \varepsilon, \tag{2}$$

where the regression error ε has the property $E(\varepsilon \mid X, v) = 0$, by construction. Windmeijer & Santos Silva (1997) consider estimation of models like (2) in situations where some of the regressors may be simultaneously determined with the dependent count. In this case, there is a crucial distinction between additive and multiplicative (for that matter structural) errors, the two otherwise being observationally equivalent (Wooldridge, 1992). Grogger (1990) discusses the additive approach and testing for exogeneity of the regressors using a Hausman-type test.

In the given context, it is natural to maintain the notation in (2) to distinguish between regression error and unobservable characteristics, the latter not being accounted for in the regression and potentially correlated with X. Mullahy (1997) gives conditions for consistent estimation of β in such a model. In a nutshell, if v and X are mean independent, then PML estimation of the Poisson model is consistent for β (Gourieroux *et al.*, 1984; Wooldridge, 1997). On the contrary, if mean independence fails, then PML will generally be inconsistent, and estimation by instrumental variables based on appropriately defined residuals is suggested alternatively. Mullahy (1997) imposes two key assumptions on the instruments Z:

$$E(v \mid Z) = E(v) \text{ and } E(Y \mid X, v, Z) = E(Y \mid X, v). \tag{3}$$

The first assumption is an independence condition that v and Z must be mean independent. The second assumption imposes an exclusion restriction on the conditional expectation function which implies for the regression error that $E(\varepsilon \mid X, Z, v) = 0$.

With the assumptions on Z, a conditional moment restriction can be constructed via the residual function $\rho(Y, X; \beta) = Y \exp(-X'\beta) - 1$ as

$$E[\rho(Y, X; \beta) \mid Z] = E[Y \exp(-X'\beta) - 1 \mid Z] = 0 \tag{4}$$

by iterated expectations. As noted by Mullahy (1997), the crucial step in deriving such a residual function is that v needs to be additively separable from X which can be achieved by dividing both sides of (2) by $\exp(X'\beta)$. The conditional moment restriction is assumed to uniquely identify the true parameter value β. Now let $a(Z)$ denote a matrix-valued function of Z with dimension $s \geq k$, which in the simplest case is the identity function $a(Z) = Z$. It is common practice to derive unconditional (population) moment restrictions from (4) as

$$E[a(Z)\rho(Y, X; \beta)] = 0, \tag{5}$$

and the estimator of β is obtained as the solution to sample analogues $\sum_i a(z_i)\rho(y_i, x_i; \hat{\beta}) = 0$, with estimation operationalized, for example, in a GMM or non-linear instrumental variables framework. Such a procedure, however, is suboptimal for at least two reasons. First, the conditional moment restriction is stronger than the unconditional ones implying that an estimator based on the latter does not necessarily exploit all the available information.

Second, the procedure is only valid under the presumption that $a(Z)$ (or in the simplest case Z) identifies β, which must not necessarily be so; see Dominguez & Lobato (2004).

In constructing the optimal instrument matrix $a^*(Z)$ both these issues need to be taken into account. More formally, let $\mathcal{D}(Z)=E[\partial\rho(Y,X;\beta)/\partial\beta' \mid Z]$ denote the Jacobian, and let $\mathcal{V}(Z)=E[\rho(Y,X;\beta)^2 \mid Z]$ denote the variance obtained from the conditional moment restriction in (4). Chamberlain (1987) shows that the asymptotic efficiency bound for any $\sqrt{n}$-consistent semiparametric estimator based on (4) is given by $\mathcal{I}^{-1}=E_Z[\mathcal{D}(Z)'\mathcal{V}(Z)^{-1}\mathcal{D}(Z)]^{-1}$. This efficiency lower bound is derived under the assumption of i.i.d. data following a multinomial distribution, in which case the usual parametric efficiency bound applies. As any distribution can be approximated arbitrarily close by the multinomial distribution, and the efficiency bound does not depend on the support of the distribution, the bound derived under the multinomial distribution also applies in the general semiparametric case.

An optimal GMM estimator based on the unconditional moment restrictions in (5) that attains the semiparametric efficiency bound requires instruments

$$a^*(Z)=\mathcal{D}(Z)'\mathcal{V}(Z)^{-1}$$

(Newey, 1993, among others). In general, such an estimator is not feasible as both expectations forming $a^*(Z)$ are unknown. It is shown in Chamberlain (1987) that a GMM estimator based on a particular sequence of unconditional moment restrictions may come arbitrarily close to the semiparametric efficiency bound. Related to this idea, Donald *et al.* (2003) use a series of functions of Z to form unconditional moment restrictions, and let the dimension K of the vector of approximating functions grow with the sample size. Let $q^K(Z)$ denote such a vector. Under relatively weak regularity conditions, mainly including that second moments exist and are finite, and that K grows sufficiently large, the sequence of unconditional moment restrictions

$$E[q^K(Z)\rho(Y,X;\beta)]=0 \tag{6}$$

is equivalent to the conditional moment restriction in (4). This is the important step to obtain unconditional moments from the model assumptions. Semiparametric efficiency is established if linear combinations of $q^K(Z)$ can approximate $a^*(Z)$, with approximation error diminishing as K grows, as the asymptotic variance of the optimal GMM estimator with instruments $a^*(Z)$ reaches the semiparametric efficiency bound (Newey, 1993).

Donald *et al.* (2003) suggest using splines as approximating functions. If Z is univariate, the sth order spline with knots $t_1,\ldots,t_{K-s-1}$ is given by

$$q^K(Z)=(1,Z,\ldots,Z^s,[1(Z>t_1)Z]^s,\ldots,[1(Z>t_{K-s-1})Z]^s)' \tag{7}$$

with indicator function $1(\cdot)$. Common choice is $s=3$ for cubic splines. For example, with three knots and cubic splines, the vector of approximating functions is

$$q^7(Z)=(1,Z,Z^2,Z^3,[1(Z>t_1)Z]^3,[1(Z>t_2)Z]^3,[1(Z>t_3)Z]^3)'$$

where the knots t_1, t_2 and t_3 could be the 0.25-quantile, the median and the 0.75-quantile of Z, respectively. For Z multivariate, the approximating functions may be generated by products of univariate splines for each element of Z. See de Boor (2001) for the theoretical background and further details. The method can be easily implemented in existing procedures that utilize unconditional moment restrictions, a potential advantage over alternative approaches such as Kitamura *et al.* (2004) and Dominguez & Lobato (2004).

3. Estimation methods and moment selection

3.1. Generalized method of moments

The GMM principle has become a well-established estimation technique for moment conditions such as (6) since Hansen (1982); see also Hall (2005). To describe it, let $g_i(\beta) = q^K(z_i)\rho(y_i, x_i; \beta)$ and $\hat{g}_n(\beta) = \sum_{i=1}^n g_i(\beta)/n$, where lower-case letters y_i, x_i, z_i denote the observed sample values of Y, X and Z. The GMM estimator $\hat{\beta}_{\text{gmm}}$ minimizes the weighted squared distance of sample and population moments, algebraically

$$\hat{\beta}_{\text{gmm}} = \arg\min_{\beta} \hat{g}_n(\beta)' \Upsilon \hat{g}_n(\beta), \tag{8}$$

where Υ is a $K \times K$ weighting matrix. For optimal GMM, the weighting matrix is chosen such that $\Upsilon = \hat{\Omega}_n(\tilde{\beta})^{-1}$ with $\hat{\Omega}_n(\beta) = \sum_{i=1}^n g_i(\beta) g_i(\beta)'/n$ and preliminary consistent estimator $\tilde{\beta}$. Under mild regularity conditions the resulting estimator $\hat{\beta}_{\text{gmm}}$ is consistent and the stabilizing transformation $\sqrt{n}(\hat{\beta}_{\text{gmm}} - \beta)$ is asymptotically normal with zero expectation and estimated covariance matrix

$$\hat{\Sigma}_{\text{gmm}} = \left[\hat{G}_n(\hat{\beta}_{\text{gmm}})' \hat{\Omega}_n(\hat{\beta}_{\text{gmm}})^{-1} \hat{G}_n(\hat{\beta}_{\text{gmm}}) \right]^{-1},$$

where $G_i(\beta) = \partial g_i(\beta)/\partial \beta'$ and $\hat{G}_n(\beta) = \sum_{i=1}^n G_i(\beta)/n$. Furthermore, the objective function scaled by $1/n$ and evaluated at the GMM estimator converges to a chi-squared distribution with $K - k$ degrees of freedom, which can be used as the basis for an overidentifying restrictions test.

To implement the approximating functions approach, one would simply use the vector of approximating functions as instrument matrix (including all other exogenous variables as well) and proceed with standard two-step GMM estimation. Clearly, the idea of using functions of the conditioning variables as additional instruments is not new; see, for example, Wooldridge (2001). In fact, one motivation of GMM is that all possible information – as given by the conditional moment restrictions – can be used in an efficient manner by choosing the 'right' weighting matrix. A general vector of approximating functions has the advantage of systematically using the information at hand, and this has not been used in this context before. Moreover, such a vector will generally improve efficiency compared with an estimator with $a(Z) = Z$, or compared with any other vague choice of $a(Z)$. On the downside, many approximating functions, and thus unconditional moment conditions, may be needed to obtain the optimal estimator.

This requirement can be a serious matter, in particular in light of recent work concerning the finite sample properties of GMM. More specifically, point estimates and inference based on the asymptotic normal distribution may be highly unreliable in finite samples and with increasing number of moments (Hansen *et al.*, 1996; Hall, 2005, among others). Alternative estimators have been proposed, for example, a bias-corrected GMM estimator in Newey & Smith (2004), and the EL estimator of Owen (1988), Qin & Lawless (1994) and Imbens (1997). Other moment estimators exist as well (Hansen *et al.*, 1996; Kitamura & Stutzer, 1997; Imbens *et al.*, 1998). Smith (1997) introduces the class of generalized EL estimators that include the aforementioned estimators as special cases, and (first-order) asymptotic equality with the GMM estimator is shown.

Further studies by Newey & Smith (2004) and Imbens & Spady (2006) examine the higher-order properties of (generalized) EL and GMM and evidence the relative advantage of the EL estimator compared with the two-step GMM estimator in terms of higher-order asymptotic bias and higher-order efficiency with increasing degree of overidentification, that is, with

increasing K. In particular, note that the optimization problem for two-step GMM implies first order conditions

$$\hat{G}_n(\hat{\beta}_{\text{gmm}})'\hat{\Omega}_n(\tilde{\beta})^{-1}\hat{g}_n(\hat{\beta}_{\text{gmm}})=0$$

and thus, in the optimum, a linear combination of sample equivalents to (6) must equal zero. It is shown, *inter alia*, that asymptotic (higher-order) bias of the two-step GMM estimator arises from estimating the Jacobian matrix (left term) and the matrix of second moments (middle term) by sample averages and the weighting matrix depending on a first step (inefficient) estimator.

As the asymptotic bias formulae are known, an analytical bias correction of $\hat{\beta}_{\text{gmm}}$ becomes available. The bias arising from estimation of the Jacobian matrix is particularly important, and a bias-corrected GMM estimator can be obtained as:

$$\hat{\beta}_{\text{bcgmm}}=\hat{\beta}_{\text{gmm}}+\hat{\Sigma}_{\text{gmm}}\sum_{i=1}^{n}\hat{G}_i\hat{P}\hat{g}_i/n, \tag{9}$$

where $\hat{g}_i=g_i(\hat{\beta}_{\text{gmm}})$, $\hat{G}_i=G_i(\hat{\beta}_{\text{gmm}})$ and $\hat{P}=\hat{\Omega}^{-1}-\hat{\Omega}^{-1}\hat{G}\hat{\Sigma}_{\text{gmm}}\hat{G}'\hat{\Omega}^{-1}$ with $\hat{G}=\hat{G}_n(\hat{\beta}_{\text{gmm}})$, $\hat{\Omega}=\hat{\Omega}_n(\hat{\beta}_{\text{gmm}})$; see Newey & Smith (2004) and Donald *et al.* (2009) for details.

In comparison with two-step GMM, other moment estimators imply first-order conditions in which the Jacobian and second moment matrix are estimated more efficiently. Among the alternatives, the EL estimator received considerable attention and was found to possess some desirable higher-order properties. In particular, it was shown that the asymptotic bias of GMM grows with the number of overidentifying restrictions, whereas the bias of EL is bounded. I will therefore discuss EL estimation of β next.

3.2. Empirical likelihood

EL estimation was first introduced in the biostatistics literature, see Owen (1988, 1991) and Qin & Lawless (1994, 1995) for details on EL and its application to moment condition models; see also Owen (2001) for a monograph on EL. More recent surveys by Imbens (2002) and Kitamura (2006) point out the richness of the EL approach, in particular as an alternative to the two-step GMM procedure.

Let p_i denote an unknown probability weight assigned to the sample outcome (y_i, x_i, z_i) of one observation i, with $0<p_i<1$ for all i, impose the normalization $\sum_i p_i=1$ and let $p=(p_1,\ldots,p_n)'$. A non-parametric likelihood estimator of p is obtained by maximizing the non-parametric log-likelihood function, algebraically

$$\hat{p}=\arg\max_{p}\sum_{i=1}^{n}\ln p_i \quad \text{such that} \quad \sum_{i=1}^{n}p_i=1. \tag{10}$$

Without further restrictions, optimal probability weights are given by $\hat{p}_i=1/n$. To incorporate special features of the data-generating process, one may impose empirical moments as additional restrictions, which can be specified from (6) as $\sum_i p_i g_i(\beta)=0$. Following Kitamura (2006), the optimization problem yields the Lagrangian function

$$\mathcal{L}=\sum_{i=1}^{n}\ln p_i+\kappa\left(1-\sum_{i=1}^{n}p_i\right)-n\lambda'\sum_{i=1}^{n}p_i g_i(\beta), \tag{11}$$

where λ and κ denote Lagrangian multipliers. It can be shown that the first-order conditions are solved by $\hat{\kappa}=n$,

$$\hat{p}_i(\beta) = \frac{1}{n[1+\hat{\lambda}(\beta)' g_i(\beta)]}, \quad \hat{\lambda}(\beta) = \arg\min_{\lambda} \left\{ -\sum_{i=1}^{n} \ln[1+\lambda' g_i(\beta)] \right\}.$$

Both optimal probability weights $\hat{p}_i$ and optimal Langrangian multipliers $\hat{\lambda}$ depend on the unknown parameter vector β. Plugging the optimality conditions into the objective function in (10) yields the empirical log-likelihood function for β

$$\ln L_{el}(\beta) = \min_{\lambda} \left\{ -\sum_{i=1}^{n} \ln\left[1+\lambda' g_i(\beta)\right] - n \ln n \right\}$$

and the EL estimator is defined as

$$\hat{\beta}_{el} = \arg\max_{\beta} \ln L_{el}(\beta) = \arg\max_{\beta} \min_{\lambda} \left\{ -\sum_{i=1}^{n} \ln\left[1+\lambda' g_i(\beta)\right] \right\}. \tag{12}$$

As maximization of (12) does not have a simple closed-form solution, numerical methods have to be applied to obtain the value of $\hat{\beta}_{el}$. Owen (2001) and Kitamura (2006) provide details on computational algorithms that have stable convergence properties in the aforementioned problem. Software codes can be found on the EL homepage related to Owen (2001); see http://www-stat.stanford.edu/∼owen/empirical/. A fast and stable code, also used in the following simulations and for the empirical example, is provided by Bruce E. Hansen (see http://www.ssc.wisc.edu/∼bhansen/progs/progs_gmm.html). He does also provide the GMM code.

Under similar regularity conditions as in the GMM framework, Qin & Lawless (1994) show the consistency of the EL estimator and prove asymptotic normality of the stabilizing transformation $\sqrt{n}(\hat{\beta}_{el} - \beta)$ with zero expectation and estimated covariance matrix

$$\hat{\Sigma}_{el} = [\hat{G}_p(\hat{\beta}_{el})' \hat{\Omega}_p(\hat{\beta}_{el})^{-1} \hat{G}_p(\hat{\beta}_{el})]^{-1},$$

where $\hat{G}_p(\beta) = \sum_{i=1}^{n} \hat{p}_i(\beta) \partial g_i(\beta)/\partial \beta'$ and $\hat{\Omega}_p(\beta) = \sum_{i=1}^{n} \hat{p}_i(\beta) g_i(\beta) g_i(\beta)'$. Note that the terms in the EL covariance matrix are estimated using probability weights $\hat{p}_i(\hat{\beta}_{el})$ obtained from an EL optimization, whereas the terms in the GMM variance are estimated using sample weights $1/n$. The EL function evaluated at the EL estimator can be used to conduct an overidentifying restrictions test as $-2[\ln L_{el}(\hat{\beta}_{el}) - (-n \ln n)]$ obeys the chi-squared distribution with $K-k$ degrees of freedom asymptotically.

It can be shown that optimal probability weights $\hat{p}_i$ and Langrangian multipliers $\hat{\lambda}$, both evaluated at the EL estimator, imply first-order conditions

$$\hat{G}_p(\hat{\beta}_{el})' \hat{\Omega}_p(\hat{\beta}_{el})^{-1} \hat{g}_n(\hat{\beta}_{el}) = 0.$$

As with two-step GMM, a linear combination of sample moments must equal zero. EL uses empirical moments for the Jacobian term and the matrix of second moments, and probability weights p_i are chosen efficiently. Moreover, the EL estimator does not depend on a preliminary, possibly inefficient estimator $\tilde{\beta}$. Based on these properties, Newey & Smith (2004) show that the EL estimator is preferable to the GMM estimator in terms of higher-order asymptotic bias, and higher-order efficiency after bias correction.

3.3. Moment selection criteria

The construction of the series expansion $q^K(Z)$ and the subsequent arguments of achieving semiparametric efficiency crucially depend on K growing large with the sample size. Under the assumption of continuously distributed instruments Z with compact support and density

bounded away from zero, Donald *et al.* (2003) derive limits on the growth rate of K. Although this provides a theoretical foundation on the choice of K, the dimension of the vector is unknown in practice for any given sample with a particular size. To construct the $q^K(Z)$ vector it would be useful to have a formal rule of how to select its dimension.

There are several ways of choosing K including cross-validation techniques (e.g. Hansen, 1982) and information criteria based on approaches known from standard likelihood methods (e.g. Andrews & Lu, 2001). I will refer to the moment selection criteria of Donald *et al.* (2009) that can be implemented in a rather straightforward manner (see also the simulation example that follows). Let $\hat{\beta}_K$ denote any of the three estimators – GMM, bias-corrected GMM or EL – given that the vector of approximating functions has dimension K. Let $t'\hat{\beta}_K$ denote a linear combination of $\hat{\beta}_K$ for some linear combination coefficients t. Let

$$\hat{\rho}_i = \rho(w_i; \hat{\beta}_K), \quad \hat{G} = \hat{G}_n(\hat{\beta}_K), \quad \hat{\Omega} = \hat{\Omega}_n(\hat{\beta}_K), \quad \hat{\Sigma} = [\hat{G}'\hat{\Omega}^{-1}\hat{G}]^{-1}, \quad \hat{\tau} = \hat{\Sigma}t,$$

$$\hat{d}_i = \hat{G}' \left[\sum_{j=1}^{n} q^K(z_j) q^K(z_j)'/n \right]^{-1} q^K(z_i), \quad \hat{\eta}_i = \partial\hat{\rho}_i/\partial\beta - \hat{d}_i,$$

$$\hat{\xi}_i = q^K(z_i)'\hat{\Omega} q^K(z_i)/n, \quad \hat{\Lambda}(K) = \sum_{i=1}^{n} (\hat{\tau}'\hat{\eta}_i)^2 \hat{\xi}_i, \quad \hat{\Pi}(K) = \sum_{i=1}^{n} (\hat{\tau}'\hat{\eta}_i)\hat{\xi}_i\hat{\rho},$$

$$\hat{\Phi}(K) = \hat{\Lambda}(K) - \hat{\tau}'\hat{\Sigma}^{-1}\hat{\tau}, \quad \hat{Q} = \sum_{i=1}^{n} q^K(z_i)\hat{\rho}(\hat{\tau}'\hat{\eta}_i) q^K(z_i)',$$

$$\hat{\Pi}_b(K) = \mathrm{tr}\left(\hat{\Omega}^{-1/2}\hat{Q}\hat{\Omega}^{-1}\hat{Q}\hat{\Omega}^{-1/2}\right), \qquad \hat{D}_i = \hat{G}'\hat{\Omega}^{-1} q^K(z_i),$$

$$\hat{\Xi}(K) = \sum_{i=1}^{n} \{5(\hat{\tau}'\hat{d}_i)^2 - \hat{\rho}^4(\hat{\tau}'\hat{D}_i)^2\}\hat{\xi}_i,$$

$$\hat{\Xi}_{el}(K) = \sum_{i=1}^{n} \{3(\hat{\tau}'\hat{d}_i)^2 - \hat{\rho}^4(\hat{\tau}'\hat{D}_i)^2\}\hat{\xi}_i.$$

The selection criteria are:

$$\begin{aligned} S_{\mathrm{gmm}}(K) &= \hat{\Pi}(K)^2/n + \hat{\Phi}(K), \\ S_{\mathrm{bcgmm}}(K) &= [\hat{\Lambda}(K) + \hat{\Pi}_b(K) + \hat{\Xi}(K)]/n + \hat{\Phi}(K), \\ S_{el}(K) &= [\hat{\Lambda}(K) - \hat{\Pi}_b(K) + \hat{\Xi}(K) - 2\hat{\Xi}_{el}(K)]/n + \hat{\Phi}(K). \end{aligned} \tag{13}$$

The optimal dimension K^* of the vector of approximating functions is chosen such that $S(K)$ is minimal, that is, $K^* = \arg\min_K S(K)$, which is shown to minimize the higher-order mean-squared error of each estimator. The terms in each criterion contain second- and higher-order moments; for details on the interpretation, see Newey & Smith (2004) and Donald *et al.* (2009).

4. Monte Carlo evidence

In this section, I compare the finite sample behaviour of EL and GMM in a generated count data experiment with correlated unobserved heterogeneity. The model imposes a conditional moment restriction as the one introduced in the previous discussion, and I investigate the performance of the proposed estimators with increasing dimension of the vector of approximating functions.

The sampling process is based on the Poisson model with Gamma-distributed heterogeneity. The model is non-standard compared with the well-known negative binomial models in

that the heterogeneity term is correlated with one of the elements in the regressor matrix X. Specifically, consider the following data-generating processes:

$$Y \mid X, v \sim \text{Poisson}(\mu_X v),$$
$$\mu_X = \exp(X'\beta), \quad v \mid \zeta \sim \text{Gamma}[1, \exp(\zeta)], \quad \zeta = \varphi + \gamma\psi - (1+\gamma^2)/2.$$

Scenario I

$$X = (1, \alpha Z + \psi)', \quad Z \sim N(0,1), \quad (\varphi, \psi) \sim \text{BVN}(0, I_2).$$

Scenario II

$$X = (1, \alpha Z + \psi, 1(\tau > 0))', \quad Z \sim \text{BVN}(0, \Sigma), \quad \Sigma = \begin{pmatrix} 1 & \sigma_{12} \\ \sigma_{12} & 1 \end{pmatrix},$$

$$(\varphi, \psi, \tau) \sim \text{TVN}(0, I_3),$$

where BVN$(\cdot)$ stands for the bivariate normal distribution, TVN$(\cdot)$ stands for the trivariate normal distribution, I_k is the k-dimensional identity matrix and $N(0,1)$ stands for the standard normal distribution. Only the triple (Y, X, Z) is observed. Scenario I is a simple design with one count-dependent variable, one endogenous continuous regressor and one instrument. Scenario II extends the setup to the more practically relevant case of mixed discrete/continuous regressors and multiple instruments to illustrate the construction of the vector of approximating functions in a multivariate setting, and to assess the role of the t vector in the selection rules (the linear combination coefficient of the parameters).

The conditional distribution of $v \mid \zeta$ is normalized such that $E(v \mid \zeta) = \exp(\zeta)$ and $\text{var}(v \mid \zeta) = \exp(2\zeta)$. The location normalization of ζ implies that $E(v) = E[E(v \mid \zeta)] = E[\exp(\zeta)] = 1$. For α fixed, the parameter γ determines the correlation between X and ζ. If γ equals zero, the unobserved heterogeneity is independent of the regressor matrix and PML consistently estimates β. For non-zero γ, the conditional expectation $E(v \mid X)$ is non-constant in X, and PML estimation will generally be inconsistent. As v and Z are independent, an assumption somewhat stronger than required, and $\alpha \neq 0$, moment estimation using Z as an instrument can be applied.

The parameter vector β is fixed at $(0,1)'$ in scenario I, and fixed at $(0,1,1)'$ in scenario II. Two different values of α are chosen (0.2 and/or 0.8) so as to vary the correlation between the instrument and regressor. Three different sample sizes are considered ($n = 100$, 500, 2000), and samples are drawn for all variables in each of 10,000 Monte Carlo replications. Calculations are carried out in Gauss v8 64-bit; EL optimization is performed using Bruce Hansen's code (see the previous link), GMM estimation is based on the built-in qnewton optimization routine (the code is available from the author upon request).

As γ is fixed at 0.5, PML estimation is inconsistent for β in each of the settings. The experiment shows that the median bias in the estimated slope varies between 0.264 and 0.381, depending on the variation in X. The results for the other estimators are displayed in Tables 1–5.

Scenario I

Consider Tables 1 and 2. For each estimation method the columns show the median bias (Med.Bias) and the median absolute deviation (MAD) of $\hat{\beta}_1$ from the true value $\beta_1 = 1$, the inter-quartile range (IQR) in the distribution of $\hat{\beta}_1$, the probability that the absolute value of the t-statistic $t = (\hat{\beta}_1 - 1)/se(\hat{\beta}_1)$ is larger than 1.96 (pVal) and the rejection rate for the over-identifying restrictions test (Over.) with 5 per cent nominal level. Robust measures of central tendency and dispersion are presented as the existence of (finite-sample) moments might be

Table 1. *Simulation results for* $\hat{\beta}_1$ *in scenario I:* $Z \sim N(0,1), \alpha = 0.2$

	GMM					BCGMM					EL				
	Med. Bias	MAD	IQR	*p*Val	Over.	Med. Bias	MAD	IQR	*p*Val	Over.	Med. Bias	MAD	IQR	*p*Val	Over.
$n=100$															
$K=2$	0.020	0.508	1.091	0.131	–	0.020	0.508	1.091	0.131	–	0.020	0.508	1.091	0.131	–
$K=4$	0.011	0.413	0.826	0.105	0.167	0.016	0.509	1.019	0.158	0.138	0.013	0.467	0.934	0.103	0.178
$K=6$	0.019	0.375	0.750	0.126	0.250	0.035	0.476	0.953	0.206	0.183	0.017	0.430	0.862	0.159	0.361
$K=8$	0.050	0.354	0.709	0.140	0.363	0.028	0.458	0.915	0.235	0.275	0.016	0.408	0.817	0.198	0.498
$K=K^*$	0.026	0.169	0.278	0.078	0.272	0.024	0.160	0.301	0.085	0.206	0.017	0.168	0.358	0.078	0.382
$n=500$															
$K=2$	0.023	0.339	0.679	0.138	–	0.023	0.339	0.679	0.138	–	0.023	0.339	0.679	0.138	–
$K=4$	0.038	0.278	0.559	0.063	0.113	0.026	0.351	0.702	0.130	0.101	0.021	0.298	0.599	0.067	0.093
$K=8$	0.046	0.229	0.459	0.094	0.199	0.047	0.316	0.631	0.211	0.133	0.012	0.290	0.580	0.168	0.253
$K=12$	0.050	0.209	0.419	0.117	0.364	0.062	0.293	0.587	0.257	0.190	0.016	0.274	0.547	0.228	0.356
$K=K^*$	0.012	0.076	0.153	0.051	0.251	0.011	0.083	0.167	0.067	0.152	0.014	0.066	0.193	0.061	0.276
$n=2000$															
$K=2$	0.014	0.189	0.381	0.022	–	0.014	0.189	0.381	0.022	–	0.014	0.189	0.381	0.022	–
$K=4$	0.022	0.173	0.348	0.048	0.084	0.008	0.200	0.400	0.086	0.082	0.008	0.178	0.359	0.060	0.071
$K=8$	0.044	0.158	0.317	0.067	0.147	0.029	0.206	0.412	0.151	0.118	0.022	0.201	0.402	0.144	0.147
$K=12$	0.060	0.147	0.294	0.078	0.196	0.049	0.207	0.414	0.198	0.118	0.017	0.207	0.414	0.206	0.199
$K=K^*$	0.015	0.053	0.108	0.041	0.164	0.010	0.057	0.113	0.050	0.126	0.010	0.045	0.131	0.051	0.169

Med.Bias is the median bias of $\hat{\beta}_1$ from the true value $\beta_1 = 1$; MAD is the median absolute deviation of $\hat{\beta}_1$ from the true value; IQR is the inter-quartile range in the distribution of $\hat{\beta}_1$; *p*Val is the probability that the absolute value of the *t*-statistic $t = (\hat{\beta}_1 - 1)/se(\hat{\beta}_1)$ is larger than 1.96; Over. is the rejection rate for an overidentifying restrictions test with 5 per cent nominal level; GMM, generalized method of moments; BCGMM, bias-corrected GMM; EL, empirical likelihood.

Table 2. *Simulation results for* $\hat{\beta}_1$ *in scenario I:* $Z \sim N(0,1), \alpha = 0.8$

	GMM					BCGMM					EL				
	Med. Bias	MAD	IQR	*p*Val	Over.	Med. Bias	MAD	IQR	*p*Val	Over.	Med. Bias	MAD	IQR	*p*Val	Over.
$n=100$															
$K=2$	0.021	0.250	0.502	0.135	–	0.021	0.250	0.502	0.135	–	0.021	0.250	0.502	0.135	–
$K=4$	−0.051	0.217	0.435	0.134	0.208	−0.034	0.231	0.463	0.149	0.186	0.013	0.209	0.419	0.121	0.266
$K=6$	−0.060	0.226	0.452	0.166	0.286	−0.033	0.256	0.511	0.201	0.225	0.005	0.219	0.440	0.156	0.446
$K=8$	−0.096	0.240	0.484	0.192	0.427	−0.053	0.282	0.563	0.248	0.338	−0.016	0.222	0.446	0.190	0.580
$K=K^*$	−0.019	0.077	0.155	0.056	0.240	−0.012	0.077	0.154	0.053	0.191	−0.010	0.074	0.180	0.048	0.389
$n=500$															
$K=2$	0.004	0.106	0.212	0.061	–	0.004	0.106	0.212	0.061	–	0.004	0.106	0.212	0.061	–
$K=4$	−0.013	0.100	0.202	0.088	0.157	−0.006	0.102	0.205	0.091	0.149	−0.001	0.101	0.202	0.093	0.175
$K=8$	−0.045	0.111	0.223	0.144	0.279	−0.027	0.122	0.245	0.166	0.215	−0.014	0.114	0.229	0.148	0.374
$K=12$	−0.083	0.128	0.258	0.196	0.472	−0.041	0.150	0.299	0.250	0.318	−0.011	0.126	0.252	0.182	0.507
$K=K^*$	−0.007	0.031	0.063	0.044	0.214	−0.004	0.030	0.060	0.043	0.174	−0.003	0.028	0.069	0.048	0.307
$n=2000$															
$K=2$	0.001	0.052	0.105	0.060	–	0.001	0.052	0.105	0.060	–	0.001	0.052	0.105	0.060	–
$K=4$	−0.008	0.050	0.101	0.071	0.110	−0.006	0.050	0.101	0.071	0.107	−0.004	0.051	0.102	0.073	0.113
$K=8$	−0.022	0.056	0.113	0.092	0.226	−0.015	0.058	0.117	0.092	0.198	−0.010	0.059	0.119	0.104	0.255
$K=12$	−0.034	0.062	0.125	0.116	0.282	−0.015	0.067	0.134	0.127	0.199	−0.008	0.067	0.134	0.127	0.331
$K=K^*$	−0.005	0.019	0.040	0.047	0.160	−0.003	0.018	0.037	0.045	0.141	−0.002	0.017	0.041	0.049	0.207

Med.Bias is the median bias of $\hat{\beta}_1$ from the true value $\beta_1 = 1$; MAD is the median absolute deviation of $\hat{\beta}_1$ from the true value; IQR is the inter-quartile range in the distribution of $\hat{\beta}_1$; *p*Val is the probability that the absolute value of the *t*-statistic $t = (\hat{\beta}_1 - 1)/se(\hat{\beta}_1)$ is larger than 1.96; Over. is the rejection rate for an overidentifying restrictions test with 5 per cent nominal level; GMM, generalized method of moments; BCGMM, bias-corrected GMM; EL, empirical likelihood.

Table 3. *Summary statistics for optimal K^* in scenario* I

	GMM				BCGMM				EL			
	Mode	1Q	2Q	3Q	Mode	1Q	2Q	3Q	Mode	1Q	2Q	3Q
$n=100$												
$\alpha=0.2$	8	5	7	9	8	5	7	8	7	4	7	8
$\alpha=0.8$	5	4	5	8	5	4	6	8	4	4	5	7
$n=500$												
$\alpha=0.2$	10	8	10	11	10	7	10	11	9	6	9	11
$\alpha=0.8$	5	4	6	9	4	4	6	9	4	4	6	9
$n=2000$												
$\alpha=0.2$	10	6	10	11	11	6	9	11	10	5	9	11
$\alpha=0.8$	5	4	5	8	5	4	6	8	4	4	6	9

$Z \sim N(0,1)$. 1Q is the first quartile, 2Q is the median and 3Q is the third quartile in the distribution of optimal K^*. GMM, generalized method of moments; BCGMM, bias-corrected GMM; EL, empirical likelihood.

an issue (e.g. Kunitomo & Matsushita, 2003; Guggenberger, 2005, 2008; Guggenberger & Hahn, 2005; Davidson & MacKinnon, 2007).

Five different specifications of $q^K(Z)$ are presented. The first, as a benchmark, is the standard instrumental variables approach with instrument Z, that is, the vector of approximating functions is simply $q^2(Z)=(1,Z)'$. The next three rows give the results with augmented instrument vector having fixed dimensions $K=4,6,8$ (for $n=100$) and $K=4,8,12$ (for $n=500$ and 2000), and the optimal dimension K^*. The approximating functions are chosen such that they form a basis for the set of cubic splines ($s=3$), and the knots $t_1,\ldots,t_{K-4}$ are set equal to the quantiles of the empirical Z-distribution. The first-step weighting matrix for the two-step GMM estimator is chosen to be the $K \times K$ identity matrix. For the selection criteria, the linear combination coefficients pick the slope as parameter of interest.

The results indicate that there are considerable efficiency gains by increasing the dimension of the vector of approximating functions. These gains are higher with a low value of α and a low number of observations. In all cases, the optimal K^* yields the lowest MAD. Owing to the variation in K^*, it is suggestive to choose the dimension of $q^K(Z)$ according to the mean-squared error criteria, as opposed to a rule-of-thumb fixed choice of K, reflected in the substantial drop in the MAD compared with the fixed K scenarios. Note that for each sample size and correlation between instrument and regressor, the median bias is lowest for optimal K^*, even lower compared with the basic specification with only Z in the vector of approximating functions. The Med.Bias and MAD tend to be smaller for EL than for the GMM estimators.

Regarding inference the results point to a size distortion for the Wald test of the hypothesis that the coefficient equals the truth. This distortion depends on K and the sample size. For the basic instrument specification ($K=2$), the p-value is larger than the nominal level only for small sample sizes, indicating a poor performance of the normal approximation. The problem becomes more persistent with a fixed $K>2$. This is likely related to the many weak instruments problem discussed in Newey & Windmeijer (2009), owing to the construction of the $q^K(Z)$ vector. The Wald size distortion is less evident, however, for the optimal choice K^*. The results do also suggest a size distortion in the overidentifying restrictions test with K growing large. Such a problem was noted, too, in the simulation study of Donald *et al.* (2009); see the comments on multiple instruments in scenario II next.

To get a feeling for the dimension of the vector of approximating functions, Table 3 displays the mode in the distribution of optimal K^*, as well as the median and the first and

Table 4. *Simulation results for* $\hat{\beta}_1$ *in scenario* II

	GMM					BCGMM					EL				
	Med.Bias	MAD	IQR	*p*Val	Over.	Med.Bias	MAD	IQR	*p*Val	Over.	Med.Bias	MAD	IQR	*p*Val	Over.
$n=100$, $\alpha=(0.2, 0.2)$															
$K=3$	0.042	0.398	0.793	0.074	0.119	0.041	0.449	0.898	0.095	0.103	0.040	0.397	0.794	0.057	0.084
$K=K^*$	0.001	0.077	0.154	0.033	0.424	0.001	0.084	0.169	0.042	0.301	0.004	0.062	0.240	0.060	0.286
$n=100$, $\alpha=(0.2, 0.8)$															
$K=3$	0.026	0.210	0.420	0.100	0.123	0.036	0.216	0.435	0.110	0.106	0.052	0.191	0.385	0.077	0.102
$K=K^*$	−0.008	0.051	0.102	0.041	0.463	−0.003	0.053	0.106	0.050	0.344	0.001	0.050	0.142	0.041	0.328
$n=100$, $\alpha=(0.8, 0.8)$															
$K=3$	0.037	0.168	0.337	0.118	0.129	0.049	0.173	0.349	0.127	0.110	0.061	0.159	0.318	0.100	0.118
$K=K^*$	−0.011	0.042	0.085	0.062	0.469	−0.006	0.043	0.086	0.041	0.348	−0.001	0.038	0.116	0.045	0.404
$n=500$, $\alpha=(0.2, 0.2)$															
$K=3$	0.037	0.222	0.445	0.053	0.055	0.025	0.240	0.481	0.072	0.051	0.011	0.206	0.412	0.040	0.057
$K=K^*$	0.008	0.051	0.087	0.041	0.391	0.009	0.045	0.096	0.031	0.213	0.004	0.046	0.098	0.048	0.502
$n=500$, $\alpha=(0.2, 0.8)$															
$K=3$	0.012	0.091	0.183	0.081	0.058	0.014	0.092	0.184	0.083	0.053	0.011	0.091	0.182	0.082	0.079
$K=K^*$	−0.004	0.021	0.046	0.045	0.317	0.002	0.027	0.051	0.039	0.217	0.002	0.017	0.049	0.052	0.533
$n=500$, $\alpha=(0.8, 0.8)$															
$K=3$	0.010	0.072	0.144	0.088	0.065	0.014	0.072	0.146	0.090	0.059	0.010	0.072	0.145	0.092	0.087
$K=K^*$	−0.002	0.017	0.039	0.032	0.325	−0.001	0.016	0.036	0.038	0.244	−0.001	0.013	0.041	0.047	0.508
$n=2000$, $\alpha=(0.2, 0.2)$															
$K=3$	0.006	0.118	0.237	0.046	0.051	−0.002	0.124	0.248	0.054	0.049	−0.008	0.117	0.234	0.047	0.056
$K=K^*$	0.009	0.031	0.033	0.045	0.151	0.005	0.035	0.070	0.042	0.091	0.006	0.032	0.085	0.041	0.128
$n=2000$, $\alpha=(0.2, 0.8)$															
$K=3$	0.003	0.047	0.094	0.060	0.050	0.003	0.047	0.094	0.060	0.050	0.001	0.047	0.095	0.062	0.066
$K=K^*$	−0.001	0.023	0.037	0.049	0.173	0.001	0.025	0.050	0.047	0.127	−0.001	0.016	0.054	0.045	0.148
$n=2000$, $\alpha=(0.8, 0.8)$															
$K=3$	0.004	0.035	0.071	0.067	0.049	0.005	0.035	0.071	0.067	0.048	0.003	0.035	0.071	0.069	0.065
$K=K^*$	−0.001	0.010	0.021	0.042	0.181	−0.001	0.011	0.023	0.041	0.154	−0.001	0.008	0.025	0.042	0.121

Med.Bias is the median bias of $\hat{\beta}_1$ from the true value $\beta_1=1$; MAD is the median absolute deviation of $\hat{\beta}_1$ from the true value; IQR is the inter-quartile range in the distribution of $\hat{\beta}_1$; *p*Val is the probability that the absolute value of the *t*-statistic $t=(\hat{\beta}_1-1)/se(\hat{\beta}_1)$ is larger than 1.96; Over. is the rejection rate for an overidentifying restrictions test with 5 per cent nominal level. The instruments are equally weighted in the moment selection criteria, that is, $t=(0, 0.5, 0.5)$. GMM, generalized method of moments; BCGMM, bias-corrected GMM; EL, empirical likelihood.

Table 5. *Mode and tendency for optimal K^* in scenario* II

	GMM	BCGMM	EL
$n=100$			
$\alpha=(0.2, 0.2)$	5, 7	5, 7	5, 7−
$\alpha=(0.2, 0.8)$	6, 5±	6, 5±	6, 5±
$\alpha=(0.8, 0.8)$	6−, 5−	6−, 5−	5, 4±
$n=500$			
$\alpha=(0.2, 0.2)$	9, 10−	9−, 10−	9, 9
$\alpha=(0.2, 0.8)$	9, 6	9, 6−	9, 6−
$\alpha=(0.8, 0.8)$	6, 6	6, 6±	6, 6
$n=2000$			
$\alpha=(0.2, 0.2)$	8+, 10	10−,10	8+, 10
$\alpha=(0.2, 0.8)$	10−, 7−	10−, 7−	10−, 7−
$\alpha=(0.8, 0.8)$	4, 5	4, 5+	4, 4

Reported numbers are mode values in the bivariate distribution of optimal approximating functions for both instruments. The first value refers to the dimension of the vector of approximating functions for the first component in Z, the second value for the second component in Z (total K is the sum of both values, plus one for the exogenous binary regressor). $\pm$ indicates that a large fraction of optimal K^* is smaller/larger than the mode and $\pm$ indicates that a large fraction of optimal K^* deviates from the mode in both directions. The instruments are equally weighted in the moment selection criteria, that is, $t=(0, 0.5, 0.5)$. GMM, generalized method of moments; BCGMM, bias-corrected GMM; EL, empirical likelihood.

third quartiles. The results are reported for each of the three estimators and for each of the settings discussed before. To achieve optimality in terms of mean-squared error, the dimension K of the vector of approximating functions is higher the lower is the correlation of instruments and regressors, and the larger is the sample size. These results are consistent with those in Donald *et al.* (2009).

Supporting Information on the journal website shows additional details for the case of log-normally distributed instruments. Although similar results are obtained regarding bias and efficiency, the Wald size distortion becomes even more evident for the log-normally distributed instruments. Regarding the dimension of the vector of approximating functions, the optimal number of elements tends to be lower than in the normal case. However, the two cases are not immediately comparable as the variation in the log-normal instrument is larger.

Scenario II

Probably more relevant from a practical point of view is the use of multiple instruments and regressors. Scenario II employs such a design. Specifically, it is assumed that one endogenous regressor can be instrumented with two normally distributed variables (with varying importance), and one additional binary variable is included in the regression model. The quantity of interest is the coefficient of the endogenous regressor, which is set to one, and the purpose of the Monte Carlo study is to show the difference between using only Z as the instrument and using the optimal vector of approximating functions, thereby fully exploiting the model assumptions. To assess the role of the t-vector in the mean-squared error criteria (the linear combination of the elements in the β-vector), two different assumptions are made. The first does equally weigh the coefficients of the endogenous and exogenous regressors, the second gives full weight on the coefficient of the endogenous regressor. The results for the former are shown in Tables 4 and 5 and the results for the latter in the Supporting Information on the journal website.

Consider Table 4 and the equal weights on coefficients. The columns provide, as before, the Med.Bias and the MAD of $\hat{\beta}_1$, the IQR, the p-value of the Wald test and the overidentifying restrictions test for each of the three estimators. Sample sizes are varied as before, and the correlations between the two instruments and the endogenous regressor are chosen as $(0.2, 0.2), (0.2, 0.8)$ and $(0.8, 0.8)$, and σ_{12} (the correlation between the two instruments) is fixed at 0.2. The first assumption on the correlation structure is closest to the following empirical application. To save on space, the results are only shown for the basic specification and the optimal vector of approximating functions. As in the simple design, the results indicate a substantial drop in the MAD by using the optimal K^*. Likewise, the Med.Bias is almost always smaller (in absolute magnitude). Overall, the three estimators perform similarly, although the EL estimator tends to have a smaller MAD than the GMM estimators. As one would expect, the MAD is lower if full weight is placed on the coefficient of the endogenous regressor.

Irrespective of the weighting scheme, the simulation suggests a size distortion in the Wald test and in the overidentifying restrictions test. Using the optimal K^* yields about the correct size of the Wald test for the small samples, for the larger samples, however, the size of the test tends to be lower than the nominal size (although not substantially). For large K, the overidentifying restrictions test rejects the null hypothesis of valid instruments too often. This might be related to a near singularity problem in the weighting matrix of GMM and in the variance of the first-order conditions of EL (Caner, 2008) owing to the construction of the vector of approximating functions. Although this seems to have an effect on the size of the overidentifying restrictions test, it does not result in a bias in any of the three estimators.

Table 5 displays the mode in the distribution of the optimal K^*s. The first value corresponds to the optimal value for the first instrument and the second corresponds to the optimal value for the second instrument. As the distribution often has substantial mass around the mode, the tendency of optimal values is indicated by a $+/-$ (for larger and smaller values), or $\pm$ if a large fraction of optimal K^* values deviates from the mode in both directions.

The results indicate that the optimal dimension of the vector of approximating functions tends to be slightly smaller than the 'univariate' optimal K^* values (Table 3). This is to be expected because of the correlation of 0.2 between the two instruments. Further results (not shown in the table) suggest that with a zero correlation, the optimal values obtained if only one instrument is used at a time are an almost perfect predictor for the optimal dimension if all instruments are used simultaneously. With a non-zero correlation, the complementarity between instruments is reflected in a smaller number of elements in the joint $q^K(Z)$-vector to obtain efficiency in terms of the mean-squared error criteria.

5. Cigarette demand and smoking habits

As a final exercise, I apply the proposed methods to the estimation of a cigarette demand function. Cigarette demand is measured as the number of cigarettes smoked per day, and thus Y has the character of a count-dependent variable. The demand for cigarettes depends on observable and unobservable characteristics and is modelled as in (1). Owing to the exponential form, the effect of a change in one regressor X_k on the expected demand can be interpreted as a semielasticity, or elasticity if the regressor is in log-form. This follows from the partial derivative of $E(Y \mid X, v; \beta)$ with respect to X_k normalized by the conditional expectation, which equals the kth element β_k in the parameter vector.

Mullahy (1985) studies the dynamic link between today's demand for cigarettes and an individual's smoking habits amassed over lifetime. If included in a regression model, such

habits can be interpreted as a lagged dependent variable, and there is good reason to believe that unobserved smoking determinants are dynamically linked as well. One would therefore suspect that, given a positive correlation between unobservables over time, the smoking habit dynamics may be overestimated in a simple Poisson regression model, and moment estimation with a suitable instrument Z as outlined before may help to avoid such problems.

The analysis is based on a subsample of $n = 1140$ male observations of the data used in Mullahy (1997); see also Mullahy (1985) for a description. The data stem from the Smoking Supplement of the 1979 US National Health Interview Survey and contain information on the respondent's socioeconomic characteristics as well as information on various health topics and smoking behaviour. For the regressions, the dependent variable has been scaled to the number of cigarette packs smoked per day (number of cigarettes divided by 20). Mullahy (1985) constructs the smoking habit measure from the total time smoked and the number of cigarettes consumed. This measure is zero for non-smokers, and positive for smokers, the exact value depending on the discount rate (here 10 per cent) and not having direct unit interpretation. Apart from the smoking habit measure as the key variable of interest, the estimated models control for age (in years), the years of schooling, a dummy variable indicating race, family income (in thousand US dollars), household size, average state-level cigarette price (in US dollars per pack in 1979) and an indicator whether smoking in restaurants had been restricted (in 1979).

The excluded instruments are the cigarette price in 1978 and the total number of years smoking in restaurants had been restricted (before and with 1979). The rationale for the instruments is that both should affect smoking habits, that is, smoking behaviour in 1978 and before, but they should not have a direct effect on current cigarette demand. The latter exclusion restriction is plausible, since cigarette prices and indicators of smoking restrictions in 1979, that is, at the time current cigarette demand is recorded, are explicitly controlled for, and thus there is no reason to believe why the instruments should have an effect on Y other than the habits channel. Compared with the data in Mullahy (1997), I restrict the sample to individuals aged 24 or younger, as those are the most responsive to changes in the instruments.

Table 6 displays the results for the smoking habit coefficient. The columns correspond to the Poisson (PML) estimator, the two-step and bias-corrected GMM estimators and the EL estimator. For the ease of exposition, the estimated parameters and standard errors have been multiplied by 100. The PML estimate shows a value of 1.253 with estimated standard error 0.081. As this is a semielastic model, the reported coefficients can be directly interpreted as approximate per cent changes in the number of packages smoked per day, that is, the value indicates a *ceteris paribus* increase by about 1.25 per cent. For the exact change, one needs to use the formula $100[\exp(1.253/100) - 1]$, which gives a 1.26 per cent change for an unit increase in the smoking habit measure. The exact change is very close to the approximate number as the coefficients are very small in absolute value. Multiplied by the average value of the smoking habits (35.65), this gives an elasticity of 0.45, that is, if the smoking habit measure increases by 1 per cent, then the expected number of cigarettes smoked per day (measured in packs) increases by 0.45 per cent. The elasticity may of course be evaluated at other values than the average smoking habits.

Using the basic instrumental variables setting with instruments all regressors except the smoking habits plus the cigarette price in 1978 and the number of years the smoking restrictions had been in place, the estimated parameters drop by around 5–10 per cent with much larger standard error. The point estimates confirm the expectation that PML might overestimate the true smoking habit effect. On the downside, from a statistical point of view, smoking habits do not significantly affect current smoking behaviour, which contradicts the

Table 6. *The effect of smoking habits on cigarette demand*

	Poisson ML	GMM	BCGMM	EL
	1.253			
	(0.081)			
Basic instruments		1.133	1.204	1.163
		(1.435)	(1.498)	(1.497)
		[0.59]	[0.58]	[0.58]
Optimized over				
(a) restaurant smoking restrictions {5}		0.805	0.723	0.717
		(0.613)	(0.589)	(0.590)
		[2.04]	[2.05]	[1.95]
(b) Cigarette price in 1978 {1}		0.142	−0.057	−0.098
		(0.727)	(0.687)	(0.690)
		[2.28]	[2.04]	[2.04]
(a) and (b) {6}		0.704	0.582	0.558
		(0.580)	(0.549)	(0.535)
		[7.70]	[7.52]	[7.92]
(a) and (b) plus interaction {7}		0.634	0.447	0.709
		(0.550)	(0.511)	(0.540)
		[7.69]	[7.73]	[8.18]
All variables		0.886	0.761	0.708
{GMM, BCGMM: 19; EL: 21}		(0.404)	(0.385)	(0.351)
		[26.91]	[24.25]	[24.88]

All models control for age, years of schooling, dummy variables indicating race and smoking restrictions in 1979, cigarette price in 1979, household income and household size. The first value is the estimated coefficient; the second value (in round brackets) is the estimated asymptotic standard error; the third value (in square brackets) is the overidentifying test statistic with degrees of freedom being the number in curly brackets +1.
Excluded instruments: Cigarette price in 1978; number of years smoking restrictions in place. In curly brackets is the number of additional elements, compared with the basic set of instruments, according to the specification of the $q^K(Z)$ vector. Optimization over all variables adds functions of the included instruments and interactions.
ML, maximum likelihood; GMM, generalized method of moments; BCGMM, bias-corrected GMM; EL, empirical likelihood.

perspective of smoking habits entering cigarette demand as a psychological and/or physiological addiction. Note that the overidentifying test statistic is sufficiently small as to not reject the null hypothesis of valid instruments. Note too that the basic setting does not fully exploit the model assumptions and, given that the instruments fulfil the mean independence assumption, an improvement over these results might be possible.

The remaining of Table 6 shows the estimation results for various specifications of the vector of approximating functions. Among the many options to specify this vector, a reasonable working guess is to first find the optimal dimension, say K_l^* for the lth element of the instrument vector, given basic specification for all other instruments, and then gradually combine the optimal K_l^* including interactions if suitable. The table first reports the results for the optimal specification of the excluded instruments, that is, the number of years smoking restrictions had been in place and the cigarette price in 1978, respectively. In curly brackets is the number of additional approximating functions, for example, for the cigarette price in 1978 its square has been additionally included. This number plus one are the degrees of freedom for the overidentifying restrictions test with test statistic reported in square brackets.

The point estimates of the smoking habit coefficient drop compared with PML and basic instrument specification. Using the square of cigarette prices in 1978 as additional instrument even turns the sign of the coefficient negative for bias-corrected GMM and EL. Although

the overidentifying restrictions are not rejected, there is only a minor gain in the value of the moment selection criteria in this case. For the restaurant smoking restrictions, the overidentifying restrictions are not rejected either, but there is a considerable drop in the value of the selection criteria, indicating higher potential efficiency gains by adding the approximating functions. Note that in both cases the null hypothesis of a zero coefficient cannot be rejected. Clearly, the element-wise optimization may be performed for the included instruments as well.

Next, I combine the optimal approximating functions for each excluded instrument to further explore the model assumptions. As suggested by the simulation study, the optimal number of approximating functions K_l^* for each instrument can be combined to obtain the optimal number of approximating functions when both instruments are considered simultaneously. Presumably, this result is specific to the data (because of the relatively low correlation between both instruments) and does not hold in general, but in any case, such a strategy is a good starting point to explore the validity of mean independence. Using the additional approximating functions and including interactions does not change the point estimates much but the standard errors become smaller owing to the additional information that is used.

Finally, combining the optimal dimension K_l^* for excluded and included instruments and adding interactions if the indicated optimal vector of approximating functions for GMM has five additional terms for restaurant smoking restrictions, household size and the cigarette price in 1979, two additional terms for family income and the square of cigarette price in 1978. For the EL estimator, the interaction between smoking restrictions and cigarette prices in 1978 and an additional term for family income has been included. The results show point estimates of 0.886 for two-step GMM, 0.761 for bias-corrected GMM and 0.708 for EL. In terms of elasticities, a 1 per cent increase in the smoking habit measure leads to an increase in the expected number of cigarette packs consumed per day by about 0.32 per cent for the GMM estimator, 0.27 per cent for the bias-corrected GMM estimator and 0.25 per cent for the EL estimator, respectively. In all cases, the estimated coefficients are statistically different from zero at the 5 per cent level.

6. Conclusion and discussion

This article extends the previous literature on instrumental variables estimation of count data models with correlated unobserved heterogeneity (e.g. Mullahy, 1997). Based on transformed residuals and a mean independence assumption, the model implies conditional moment restrictions that can be estimated by common moment estimators such as GMM and EL. As the asymptotic variance typically depends on the choice of instruments, the article proposes the use of a general vector of approximating functions, opting ideas of Donald *et al.* (2003), to improve the efficiency of the resulting estimator. The benefits of this approach are demonstrated in a Monte Carlo study and an empirical example of a cigarette demand function.

Overall, the EL estimator tends to perform better than two-step GMM and the bias-corrected GMM estimator in terms of bias and efficiency, although the three estimators do not differ substantially in the considered scenarios. As the EL objective function includes an inner and outer loop optimization, current EL optimization routines are significantly slower in computation times compared with existing GMM routines. Furthermore, convergence tends to be harder to achieve in the case of the EL estimator (in the aforesaid simulations convergence failed in about 5–10 per cent of the replications). In terms of computational burden, one might thus prefer the simpler two-step GMM estimator or bias-corrected GMM estimator.

The simulations indicate that further research is needed regarding inference. First, the variance correction proposed in Newey & Windmeijer (2009) for many weak instruments might be assessed for the fixed K scenarios. Second, the performance of various testing procedures for the overidentifying restrictions test need to be compared and whether they are insensitive to the construction of the vector of approximating functions. Finally, one might extend the analysis of the inference issues with respect to the whole class of generalized EL estimators.

Acknowledgements

The author thanks the editor, an associate editor, two anonymous referees, as well as Guido Imbens, Joao Santos Silva, Richard Smith, Rainer Winkelmann, Tiemen Woutersen, Frank Windmeijer and participants of meetings in Zurich, Dresden and Lisbon for their valuable comments. John Mullahy kindly provided the data for the empirical part. This is a substantially revised version of SOI Working Paper No. 0404 'Empirical Likelihood in Count Data Models: The Case of Endogenous Regressors' and SOI Working Paper No. 0704 'Count Data Models with Unobserved Heterogeneity: An Empirical Likelihood Approach'.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1. Summary statistics for optimal K^* in scenario I.
Table S2. Simulation results for $\hat{\beta}_1$ in scenario I: $Z \sim LN(0,1), \alpha = 0.2$.
Table S3. Simulation results for $\hat{\beta}_1$ in scenario I: $Z \sim LN(0,1), \alpha = 0.8$.
Table S4. Simulation results for $\hat{\beta}_1$ in scenario II: $t = (0,1,0)$.
Table S5. Mode and tendency for optimal K^* in scenario II.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

References

Andrews, D. W. K. & Lu, B. (2001). Consistent model and moment selection procedures for GMM estimation with application to dynamic panel data models. *J. Econometrics* **101**, 123–164.
de Boor, C. (2001). *A practical guide to splines*. Revised edition. Springer-Verlag, New York.
Caner, M. (2008). Nearly-singular design in GMM and generalized empirical likelihood estimators. *J. Econometrics* **144**, 511–523.
Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *J. Econometrics* **34**, 305–334.
Cheung, Y. B. (2002). Zero-inflated models for regression analysis of count data: a study of growth and development. *Stat. Med.* **21**, 1461–1469.
Davidson, R. & MacKinnon, J. G. (2007). Moments of IV and JIVE estimators. *Econom. J.* **10**, 541–553.
Delgado, M. A. & Kniesner, T. J. (1997). Count data models with variance of unknown form: an application to a hedonic model of worker absenteeism. *Rev. Econ. Stat.* **79**, 41–49.
Dominguez, M. & Lobato, I. (2004). Consistent estimation of models defined by conditional moment restrictions. *Econometrica* **72**, 1601–1615.
Donald, S. G., Imbens, G. W. & Newey, W. K. (2003). Empirical likelihood estimation and consistent tests with conditional moment restrictions. *J. Econometrics* **117**, 55–93.
Donald, S. G., Imbens, G. W. & Newey, W. K. (2009). Choosing instrumental variables in conditional moment restriction models. *J. Econometrics* **152**, 28–36.
Gourieroux, C., Monfort, A. & Trognon, A. (1984). Pseudo maximum likelihood methods: applications to Poisson models. *Econometrica* **52**, 701–720.
Grogger, J. T. (1990). A simple test for exogeneity in probit, logit, and Poisson regression models. *Econom. Lett.* **33**, 329–332.

Guggenberger, P. (2005). Monte-Carlo evidence suggesting a no moment problem of the continuous updating estimator. *Econom. Bull.* **3**, 1–6.

Guggenberger, P. (2008). Finite sample evidence suggesting a heavy tail problem of the generalized empirical likelihood estimator. *Econometric Rev.* **27**, 526–541.

Guggenberger, P. & Hahn, J. (2005). Finite sample properties of the two-step empirical likelihood estimator. *Econometric Rev.* **24**, 247–263.

Gurmu, S., Rilstone, P. & Stern, S. (1998). Semiparametric estimation of count regression models. *J. Econometrics* **88**, 123–150.

Hall, A. R. (2005). *Generalized method of moments*. Oxford University Press, Oxford.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50**, 1029–1054.

Hansen, L. P., Heaton, J. & Yaron, A. (1996). Finite-sample properties of some alternative GMM estimators. *J. Bus. Econom. Statist.* **14**, 262–280.

Hausman, J., Hall, B. H. & Griliches, Z. (1984). Econometric models for count data with an application to the patents – R&D relationship. *Econometrica* **52**, 909–938.

Imbens, G. W. (1997). One-step estimators for over-identified generalized method of moment models. *Rev. Econom. Stud.* **64**, 359–383.

Imbens, G. W. (2002). Generalized method of moments and empirical likelihood. *J. Bus. Econom. Statist.* **20**, 493–506.

Imbens, G. W. & Spady, R. H. (2006). The performance of empirical likelihood and its generalizations. In *Identification and inference for econometric models: essays in honour of Thomas Rothenberg* (ed. D. W. K. Andrews), 216–244. Cambridge University Press, Cambridge.

Imbens, G. W., Spady, R. H. & Johnson, P. (1998). Information theoretic approaches to inference in moment condition models. *Econometrica* **66**, 333–357.

King, G. (1987). Presidential appointments to the supreme court: adding systematic explanation to probabilistic description. *Am. Polit. Q.* **15**, 373–386.

Kitamura, Y. (2006). Empirical likelihood methods in econometrics: theory and practice. In *Advances in economics and econometrics: theory and applications. Ninth World Congress, Vol. 2* (eds R. Blundell, W. K. Newey & T. Persson), 174–237. Cambridge University Press, Cambridge.

Kitamura, Y. & Stutzer, M. (1997). An information-theoretic alternative to generalized method of moments estimation. *Econometrica* **65**, 861–874.

Kitamura, Y., Tripathi, G. & Ahn, H. (2004). Empirical likelihood based inference in conditional moment restriction models. *Econometrica* **72**, 1667–1714.

Kunitomo, N. & Matsushita, Y. (2003). Finite sample distributions of the empirical likelihood estimator and the GMM estimator. CIRJE Discussion Paper F-200. Available at http://www.e.u-tokyo.ac.jp/cirje/research/dp/2003/2003cf200.pdf (accessed February 14, 2010).

Mullahy, J. (1985). Cigarette smoking: habits, health concern, and heterogeneous unobservables in a microeconometric analysis of consumer demand. PhD Dissertation. University of Virginia.

Mullahy, J. (1997). Instrumental-variable estimation of count data models: applications to models of cigarette smoking behavior. *Rev. Econ. Stat.* **79**, 586–593.

Newey, W. K. (1993). Efficient estimation of models with conditional moment restrictions. In *Handbook of statistics*, Vol. 11 (eds G. Maddala, C. Rao & H. Vinod), chapter 16. Elsevier Science, North Holland.

Newey, W. K. & Smith, R. J. (2004). Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica* **72**, 219–255.

Newey, W. K. & Windmeijer, F. A. G. (2009). Generalized method of moments with many weak moment conditions. *Econometrica* **77**, 687–719.

Owen, A. B. (1988). Empirical likelihood ratio confidence regions for a single functional. *Biometrika* **75**, 237–249.

Owen, A. B. (1991). Empirical likelihood for linear models. *Ann. Statist.* **19**, 1725–1747.

Owen, A. B. (2001). *Empirical likelihood*. Chapman & Hall/CRC, Boca Raton.

Pohlmeier, W. & Ulrich, V. (1995). An econometric model of the two-part decisionmaking process in the demand for health care. *J. Hum. Resour.* **30**, 339–361.

Qin, J. & Lawless, J. (1994). Empirical likelihood and general estimating equations. *Ann. Statist.* **22**, 300–325.

Qin, J. & Lawless, J. (1995). Estimating equations, empirical likelihood, and constraints on parameters. *Canad. J. Statist.* **23**, 145–159.

Santos Silva, J. M. C. & Tenreyro, S. (2006). The log of gravity. *Rev. Econ. Stat.* **88**, 641–658.

Smith, R. J. (1997). Alternative semi-parametric likelihood approaches to generalised method of moments estimation. *Econ. J.* **107**, 503–519.
Thall, P. F. & Vail, S. C. (1990). Some covariance models for longitudinal count data with overdispersion. *Biometrics* **46**, 657–671.
Windmeijer, F. A. G. & Santos Silva, J. M. C. (1997). Endogeneity in count data models: an application to demand for health care. *J. Appl. Econometrics* **12**, 281–294.
Winkelmann, R. (2008). *Econometric analysis of count data*, 5th edn. Springer-Verlag, Berlin Heidelberg.
Winkelmann, R. & Zimmermann, K. F. (1994). Count data models for demographic data. *Math. Popul. Stud.* **4**, 205–221.
Wooldridge, J. M. (1992). Some alternatives to the Box–Cox regression. *Internat. Econom. Rev.* **33**, 935–955.
Wooldridge, J. M. (1997). Quasi-likelihood methods for count data. In *Handbook of applied econometrics Vol. 2 – microeconomics* (eds M. H. Pesaran & P. Schmidt), 352–406. Blackwell Publishers, Oxford.
Wooldridge, J. M. (2001). Applications of generalized method of moments estimation. *J. Econ. Perspect.* **15**, 87–100.

Received January 2008, in final form October 2009

Stefan Boes, Socioeconomic Institute, University of Zurich, Zuerichbergstrasse 14, CH-8032 Zurich, Switzerland.
E-mail: boes@sts.uzh.ch